

Xây Dựng Và Đánh Giá

Bộ Dữ Liệu Nhận Dạng Món Ăn Việt Nam

Trần Quang Linh^{1,2,*}, Lâm Gia Huy^{1,2,*}, Trần Trung Hiếu^{1,2,*},
Lê Quang Nhật^{1,2,*}, Nguyễn Văn Kiệt^{1,2,†}

¹Trường Đại Học Công Nghệ Thông Tin

²Đại Học Quốc Gia Thành Phố Hồ Chí Minh

Email: ^{*}{18520997, 18520832, 18520754, 18521190}@gm.uit.edu.vn,
[†]kietnv@uit.edu.vn

Tóm tắt nội dung Trong bài báo này, chúng tôi tiến hành xây dựng và phát triển bộ dữ liệu về hình ảnh các món ăn Việt Nam với tổng cộng 12017 ảnh thuộc 15 món ăn. Bên cạnh đó, chúng tôi đã áp dụng phương pháp học máy truyền thống cùng các phương pháp học sâu để chạy thực nghiệm trên bộ dữ liệu này, từ đó đánh giá mức thích hợp của bộ dữ liệu đối với bài toán phân lớp ảnh. Chúng tôi đã chọn ba phương pháp để thực nghiệm gồm một phương pháp truyền thống là Logistic Regression và hai phương pháp học sâu là MobileNetV2 và DenseNet121. Sau khi hoàn thành chạy thực nghiệm và đánh giá, mô hình DenseNet121 cho kết quả cao nhất với F1-score là 82% và Top-5 Accuracy là 98%. Chúng tôi mong muốn ứng dụng kết quả nghiên cứu vào thực tế để nâng cao việc quảng bá món ăn truyền thống Việt Nam tới du khách cũng như văn hóa ẩm thực Việt Nam tới thế giới.

Keywords: Dữ Liệu Món Ăn - Máy Học - Học Sâu - Phân Loại

1 Giới thiệu

Với sự phát triển ngày càng mạnh mẽ của Công nghệ Thông tin, các công cụ ứng dụng Trí tuệ nhân tạo (AI) đóng một vai trò đặc biệt quan trọng trong việc giải quyết các bài toán nhận dạng dạng vật thể. Các bài toán đã được giải quyết và đạt được hiệu quả ứng dụng rất cao trong thực tế như "Nhận diện khuôn mặt" [1], "Nhận diện biển báo giao thông" [14]. Lấy cảm hứng từ đó, nhóm chúng tôi đã xây dựng một bộ dữ liệu về các món ăn truyền thống của Việt Nam. Chúng tôi sẽ dựa vào các phương pháp học máy và học sâu để thực hiện quá trình phân loại hình ảnh của các món ăn. Cuối cùng sau khi hoàn thành quá trình phân loại, chúng tôi sẽ sử dụng mô hình phân loại này để phục vụ cho việc tìm kiếm thông tin về các món ăn. Thông qua hình ảnh do người dùng chụp lại, mô hình sẽ dự đoán rằng đó là món ăn nào và đưa ra thông tin cụ thể hơn về món ăn đó. Với ứng dụng trên, chúng tôi có thể quảng bá các món ăn truyền thống của Việt Nam đến các bạn bè du khách quốc tế cũng như những người không có nhiều kiến thức về ẩm thực Việt Nam. Từ đó họ có thể tìm hiểu sâu hơn về các món ăn mà họ yêu thích thông qua hình ảnh.



Hình 1: Bún bò

Ví dụ như Hình 1 ở trên, đây là món bún bò và nó có nhiều đặc điểm giống với các món khác như Phở, Bún giò, v.v. Chính vì vấn đề này mà chúng tôi muốn xây dựng một mô hình để phân loại ảnh các món ăn. Như vậy, đầu vào của bài toán này sẽ là một ảnh của một món ăn và đầu ra của bài toán sẽ là tên món ăn và thông tin chung của món ăn đó.

Với phương pháp học máy truyền thống, chúng tôi sẽ sử dụng Logistic Regression để kiểm tra hiệu suất chung của mô hình. Từ đó điều chỉnh dữ liệu một cách thích hợp hơn. Đối với các phương pháp sâu như MobileNetV2 và DenseNet121, chúng tôi sẽ thực hiện quá trình đánh giá một cách chi tiết hơn để xem kết quả của quá trình phân loại đạt được ở mức nào. Sau khi trải qua quá trình thực nghiệm, chúng tôi nhận thấy rằng các phương pháp học sâu của chúng tôi đã giải quyết tương đối tốt nhiệm vụ phân loại hình ảnh của các món ăn.

Ở các mục tiếp theo sẽ là từng bước trong quá trình mà chúng tôi thực hiện. Ở mục 2 sẽ là các công trình liên quan đến dữ liệu về món ăn cũng như các phương pháp xây dựng dữ liệu. Mục 3 sẽ nêu chi tiết về bộ dữ liệu từ cách thu thập, gán nhãn đến kiểm tra. Các phương pháp tiếp cận sẽ được giới thiệu trong mục 4 của bài báo này. Mục 5 sẽ đề cập đến các thực nghiệm và thảo luận. Cuối cùng là mục 6 kết luận và hướng phát triển.

2 Các công trình nghiên cứu liên quan

Trong phần này, chúng tôi giới thiệu các công trình xây dựng bộ dữ liệu món ăn trên ở một số nước.

Parneet Kau, Karan Sikka, Weijun Wang, Serge Belongie và Ajay Divakaran (FoodX-251: A Dataset for Fine-grained Food Classification 2019) [6] là bài báo về xây dựng bộ dữ liệu phân loại món ăn. Với 118.000 ảnh cho tập huấn luyện (nhận được phân loại bằng từ loại của ảnh khi tải về), 12.000 ảnh cho tập kiểm thử và 28.000 ảnh cho tập kiểm tra (cả hai tập trên được gán nhãn bởi con người). Bộ dữ liệu được chạy thử nghiệm bằng phương pháp máy học ResNet-101(all-layers) và đạt được kết quả 17% dựa theo độ đo top-3 error rate trên từng tập dữ liệu.

Xin Chen, Yu Zhu, Hua Zhou, Liang Diao và Dongyan Wang (ChineseFoodNet: A Large-scale Image Dataset for Chinese Food Recognition) Xây dựng bộ dữ liệu món ăn

của Trung Quốc [2]. Cụ thể, bộ dữ liệu bao gồm 185,628 ảnh với 208 nhãn ứng với các món ăn truyền thống của Trung Quốc được thu thập từ mạng xã hội Douguo¹. Các nhà nghiên cứu của bài báo này đã sử dụng nhiều phương pháp học sâu (CNNs) như (ResNet, DenseNet, VGG19-BN, SqueezeNet1) để xây dựng mô hình.

Một số bài báo khác liên quan đến việc xây dựng bộ dữ liệu món ăn cũng được chúng tôi liệt kê trong bảng 1.

Bảng 1: Các bộ dữ liệu về món ăn đã được xây dựng

Bộ dữ liệu	Số lớp	Tổng số ảnh	Nguồn	Loại món ăn
ChineseFoodNet[2]	208	185,628	Web	Trung Quốc
NutriNet dataset[9]	520	225,953	Web	Nội Châu Âu
Food-251[6]	251	158,846	Web	Nhiều loại
VNFOOD-15	15	12,017	Web	Món ăn Việt

Từ các công trình trên thế giới, chúng tôi nhận thấy cần phải xây dựng một bộ dữ liệu về món ăn ở Việt Nam vì ẩm thực Việt Nam rất đa dạng và phong phú, nhiều món ăn khó phân biệt.

3 Bộ dữ liệu

Trong phần này, chúng tôi trình bày các thông tin cơ bản về bộ dữ liệu, quy trình thu thập và các thách thức mà chúng tôi phải đối mặt trên bộ dữ liệu các món ăn Việt Nam - VNFOOD-15.

3.1 Thu thập dữ liệu

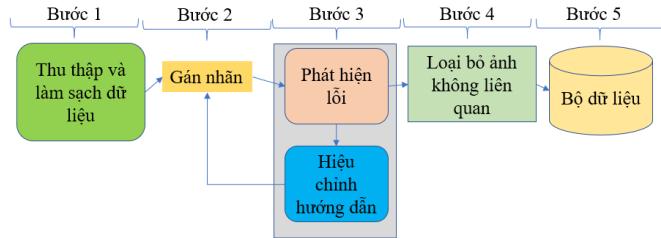
Chúng tôi thu thập dữ liệu từ các trang tìm kiếm lớn, có chứa nguồn hình ảnh với số lượng lớn và có tính đa dạng cao như Bing, Google, Instagram và Flickr. Chúng tôi sử dụng công cụ tải ảnh để tải về tự động số lượng lớn ảnh theo món ăn, sau đó lưu trữ trong các thư mục và đặt tên theo từng món ăn đó. Tiếp đến, chúng tôi sử dụng công cụ lọc ảnh trùng lặp là Duplicate Clearner Pro để lọc ảnh trùng giữa các món ăn trên tất cả các nguồn đã tải.

3.2 Kiểm tra nhãn dữ liệu

Ở bước kiểm tra nhãn dữ liệu này, mỗi người kiểm tra nhãn sẽ chịu trách nhiệm cho việc xây dựng và cập nhật hướng dẫn kiểm tra nhãn (guideline) cho ba món ăn khác nhau. Sau đó, mỗi thành viên trong nhóm kiểm tra nhãn sẽ kiểm tra nhãn chéo cho các món ăn của những thành viên khác sử dụng công cụ tạo bảng - Google Sheet. Đối với nhãn của từng ảnh, người gán sẽ gán 1 nếu ảnh đó tương ứng với đặc điểm, yêu cầu của món ăn đã được mô tả trong guideline, và gán 0 nếu ngược lại. Với mỗi thư mục ảnh (chứa một món ăn), có 3 người kiểm tra nhãn và chia thành 5 giai đoạn để kiểm

¹ www.douguo.com

tra. Sau đó, tính độ đồng thuận trung bình của từng món cho từng giai đoạn. Qua mỗi giai đoạn, nếu độ đồng thuận không đạt mức yêu cầu (chúng tôi sử dụng ngưỡng độ đồng thuận là 0.8) thì người kiểm tra nhãn sẽ được học lại guideline và bổ sung những trường hợp nhập nhằng để hoàn thiện guideline. Cuối cùng, tất cả các món ăn được kiểm tra nhãn đều có đồng thuận đạt yêu cầu (độ đồng thuận ≥ 0.8).



Hình 2: Quy trình xây dựng bộ dữ liệu

3.3 Thách thức của bộ dữ liệu

Trong quá trình xây dựng dữ liệu không tránh khỏi gặp phải những khó khăn, thách thức. Chúng tôi liệt kê những thách thức mà nhóm đã gặp phải như sau:

- Một số nguồn truy cập dữ liệu có thời gian sử dụng ngắn. Do đó, sau một khoảng thời gian nhất định thì ảnh sẽ không còn nữa và việc tải ảnh hoặc kiểm chứng ảnh trở nên khó khăn, hầu như là không thể.
- Một số món ăn được biến tấu phong phú, đa dạng về hình thức. Biến tấu về nguyên liệu do vùng miền hoặc khác biệt trong cách thức chế biến ở một số địa phương, hay khác biệt do sự du nhập văn hóa. Dẫn đến một sự khó khăn nhất định trong việc nhận dạng, kiểm tra nhãn và thống nhất guideline.
- Một số ảnh không thống nhất một định dạng tập tin chung nên khi tiến hành tiền xử lý có thể gây ra lỗi, vì vậy chúng tôi chấp nhận loại bỏ những ảnh đó;

3.4 Thông tin bộ dữ liệu

Bộ dữ liệu hoàn chỉnh bao gồm 12017 ảnh được chứa trong 15 thư mục đặt tên theo tên từng món ăn. Chúng tôi thống kê số lượng dữ liệu theo nguồn thu thập (Google, Bing, Instagram, Flickr) cho tập huấn luyện và tập kiểm tra như bảng 2 bên dưới:

Bảng 2: Thống kê tập dữ liệu

Tên Nhãn	Tập huấn luyện	Tập kiểm tra	Tổng ảnh
Xôi mặn	939	200	1139
Miến	695	200	895
Bắp xào	900	200	1100
Bún đậu	472	200	672
Phở	641	200	841
Cơm tấm	844	200	1044
Bánh canh chả cá	473	200	673
Chả giò	440	200	640
Bún bò	406	200	606
Bánh bao	519	200	719
Bánh bèo	569	200	769
Bánh mì	697	200	897
Bánh trung thu	481	200	681
Bánh xèo	407	200	607
Mì quảng	534	200	734
15 món	9,017	3,000	12,017

4 Phương pháp tiếp cận

4.1 Trích xuất đặc trưng

Histogram of Oriented Gradients - HOG: Là phương pháp biến đổi ảnh về dạng Histogram, được sử dụng trong một số bài toán phân lớp, tiêu biểu là Human Detection[3]. HOG bắt đầu từ bước chuẩn hóa ảnh và tính toán Gradients sau đó đưa ra trọng số cho các Cell thông qua số lượt bình chọn (Vote), tiếp đến là chuẩn hóa các khối (Block) và cuối cùng là xây dựng Histogram hoàn chỉnh. Chúng tôi sử dụng phương pháp HOG như là một bước xử lý kĩ thuật trước khi đưa vào mô hình tính toán.

Scale Invariant Feature Transform - SIFT[8]: SIFT là phương pháp xây dựng một đồ thị các điểm (keypoint) khoanh vùng các đặc trưng quan trọng của ảnh, và một vector mô tả (descriptor) có nhiệm vụ đại diện, mang tính phân biệt cho keypoint của ảnh đó - do đó thường được sử dụng trong các bài toán phân lớp, so sánh ảnh. Chúng tôi sử dụng phương pháp SIFT kết hợp với thuật toán phân cụm K-Means (K-Means Clustering) để xây dựng đầu vào cho mô hình học máy truyền thống.

VGG-16 [12]: Đây là thuật toán học sâu được giới thiệu vào năm 2015. VGG-16 là một kiến trúc mạng nơ-ron tích chập được đề xuất bởi Karen Simonyan , Andrew Zisserman của đại học Oxford. Chúng tôi sử dụng mô hình pre-trained VGG-16 từ thư viện Keras để trích xuất đặc trưng ảnh. Đây là một cách hiệu quả để nâng cao hiệu suất của mô hình học máy truyền thống.

4.2 Mô hình truyền thống

Logistic Regression [13]: Đây là phương pháp phân loại nhị phân, được sử dụng như là một phương pháp phân lớp cổ điển. Chúng tôi đã áp dụng mô hình Logistic Regression lên bộ dữ liệu VNFOOD-15 và đã thu được một số kết quả sơ bộ ban đầu, chuẩn bị cho những bước tiếp theo.

4.3 Mô hình học sâu

MobileNetV2 [10]: là mô hình ra đời dựa trên MobileNetV1 [4], với mong muốn giảm thiểu chi phí tính toán nhưng vẫn đạt được độ chính xác mong muốn. Nhờ vào cấu trúc phân chia theo chiều sâu (Depthwise Separable Convolution) mà số lượng tham số cũng như thời gian huấn luyện của mô hình được giảm đi đáng kể, đổi lại độ chính xác bị giảm nhẹ. Tuy nhiên, mô hình vẫn thích hợp trong việc ứng dụng vào các bài toán nhận diện vật thể (Object Detection), phân tích ngữ nghĩa (Semantic Segmentation) hay phân lớp ảnh (Image Classification).

DenseNet121 [5]: là một kiến trúc mạng nơ-ron hiện đại, được thiết kế gồm các khối (Dense Block) chứa các kết nối dày đặc (Dense Connectivity), giữa các khối là các lớp chuyển tiếp (Transition Layers) có chức năng giảm kích thước và độ sâu của kiến trúc mạng bằng Convolutional và Pooling mà không làm giảm độ chính xác của toàn mô hình. Sử dụng lớp thắt cổ chai 1x1 (Bottleneck Layer) trước mỗi lớp Convolutional 3x3 giúp giảm bớt chi phí tính toán. DenseNet121 cho ra kết quả khá tốt trên bộ dữ liệu VNFOOD-15.

5 Thực nghiệm và kết quả

5.1 Tiền xử lý dữ liệu

Mỗi một bức ảnh có rất nhiều đặc trưng khác nhau. Do đó, để có thể đưa vào mô hình sử dụng, dữ liệu ảnh phải qua một vài bước tiền xử lý. Dưới đây là sơ bộ các bước tiền xử lý trên bộ dữ liệu VNFOOD-15:

- Đọc ảnh, chuyển đổi kênh màu;
Đọc ảnh, sau đó chuyển kênh màu của tất cả các ảnh về định dạng RGB để tạo sự thống nhất trong số lượng kênh màu tất cả các ảnh và để phù hợp cho đầu vào của mô hình.
- Chuyển đổi kích thước ảnh;
Thay đổi kích thước ảnh về kích thước phù hợp - chiều cao: 224 pixel và chiều rộng: 224 pixel. Như vậy tất cả các ảnh đã được chuyển về kích thước 224*224*3.
- Chuẩn hóa ảnh;
Để đẩy nhanh tốc độ huấn luyện mô hình cũng như giảm chi phí tính toán thì mỗi giá trị pixel của ảnh sẽ được chia cho 255. Đây là kiểu chuẩn hóa MinMax vì giá trị lớn nhất mà một pixel có thể nhận được là 255 nên miền giá trị của pixel sẽ nằm trong khoảng [0, 1].
- Label Encoding (sử dụng cho MobileNetV2 và DenseNet121);
Sử dụng đối tượng LabelEncoder trong module preprocessing của thư viện sklearn để chuyển đổi nhãn từ dạng kí tự sang dạng số, sau đó chuyển về dạng OneHot, phục vụ cho việc huấn luyện các mô hình học sâu.

Sau tiền xử lý, dữ liệu đã được đưa về dạng thích hợp để huấn luyện mô hình. Chúng tôi chạy thực nghiệm trên một mô hình cổ điển là Logistic Regression và hai mô hình học sâu là MobileNetV2 và DenseNet121. Tập kiểm tra có 200 ảnh cho mỗi món. Số ảnh còn lại được chia cho tập huấn luyện và kiểm thử với tỉ lệ 8:2. Chúng tôi sử dụng K-fold Cross Validation với K = 5, test_size = 0.2 để chia dữ liệu cho tập huấn luyện và tập kiểm thử.

5.2 Thông số mô hình

Các mô hình cổ điển và học sâu đã được chúng tôi sử dụng để huấn luyện và so sánh kết quả trên bộ dữ liệu VNFOOD-15 với những bộ thông số khác nhau. Dưới đây là các thông số đã được chúng tôi cài đặt:

- **Logistic Regression - LR:** Thông số sử dụng: penalty='l2', C=1.0, random_state=0, solver='lbfgs'
- **MobileNetV2:** Chúng tôi sử dụng mô hình MobileNetV2 Pre-trained². Các thông số được chúng tôi sử dụng là: Batch_size=64, Epochs=15, Adam Optimizer với learning_rate=0.001, loss='binary_crossentropy', sử dụng metrics là 'accuracy' và TopKCategoricalAccuracy (với k=5).
- **DenseNet121** Sử dụng mô hình DenseNet121 Pre-trained³. Các thông số được chúng tôi sử dụng là: Batch_size=64, Epochs=15, Adam Optimizer với learning_rate=0.001, loss='categorical_crossentropy', sử dụng metrics là 'accuracy' và TopKCategoricalAccuracy (với k=5).

5.3 Kết quả

Trong phần này, chúng tôi trình bày về kết quả của tất cả các mô hình nêu trên sau quá trình chạy thực nghiệm.

Chúng tôi sử dụng các loại thang đo là F1-score[7], Accuracy và Top 5 accuracy[11] để đánh giá hiệu suất của các mô hình đã thử nghiệm vì có sự chênh lệch trên số lượng ảnh của mỗi lớp đối với bộ dữ liệu VNFOOD-15 mà chúng tôi đã xây dựng và sự chênh lệch về số lượng ảnh được phân loại đúng và ảnh phân loại sai.

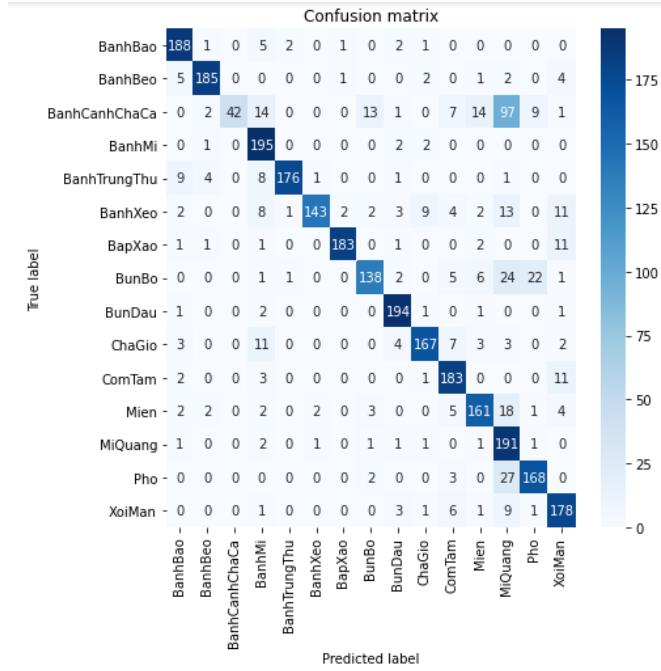
Dưới đây là kết quả đánh giá trên các mô hình và confusion matrix của mô hình DenseNet121.

Bảng 3: Kết quả đánh giá trên tập dữ liệu kiểm tra

Mô hình	F1-score	Accuracy	Top-5 accuracy
HOG + Logistic Regression	0.2048	0.2113	0.5956
VGG16 + Logistic Regression	0.7021	0.7135	0.9558
SIFT + Logistic Regression	0.3489	0.3606	0.756
MobileNetV2	0.6674	0.6703	0.9183
Densenet121	0.8227	0.8306	0.9826

² <https://keras.io/api/applications/mobilenet/>

³ <https://keras.io/api/applications/densenet/densenet121-function>



Hình 3: Confusion Matrix trên DenseNet121

Từ các kết quả ở Hình 2 và Bảng 3 chúng tôi nhận thấy, đối với mô hình học sâu càng hiện đại như DenseNet121 thì kết quả dự đoán sẽ chính xác hơn so với mô hình truyền thống như Logistic Regression. Cùng với đó là mô hình học sâu như VGG-16 kết hợp với Logistic Regression cho kết quả cao nhõ vào các đặc trưng mà VGG-16 trích xuất được. Sau đây là một vài nhận xét mà chúng tôi đưa ra sau khi đánh giá kết quả:

- Số lượng nhãn dự đoán đúng trên mô hình Hog + Logistic Regression rất ít, chỉ ở mức xấp xỉ 21%.
- Nhãn **Bánh canh chả cá** có sự sai sót nhiều (đến 158 ảnh) khi phân loại với các nhãn **Mì quảng**, **Phở** trên mô hình DenseNet121, các nhãn khác vẫn xảy ra sự sai sót nhưng số lượng rất ít (ít hơn 10 ảnh).
- Đa số các nhãn được dự đoán tốt (có độ chính xác cao) trên DenseNet121, trong đó đặc biệt là các món **Bánh Mì**, **Bún Đậu**.

5.4 Phân tích lỗi

Sự phân loại sai của mô hình đa phần do tỷ lệ tương đồng cao trong đặc điểm, thành phần cấu tạo của món ăn, dẫn đến rất khó để có thể phân biệt được. Một số ví dụ về hình ảnh khó phân biệt, có thể gây nhầm lẫn.



(a) Món Bánh canh chả cá dự đoán sai (nhầm Bún bò)
(b) Ảnh món Bún bò dự đoán sai (nhầm với món Phở)
(c) Ảnh món Bánh mì dự đoán sai (nhầm với món Chả giò)
(d) Ảnh món Xôi mặn dự đoán sai (nhầm với món Cơm tấm)

Hình 4: Một số trường hợp điển hình cho việc dự đoán sai của mô hình

Việc nhận diện món ăn thông qua một khía cạnh của bức ảnh là không hề dễ dàng vì sự tương đồng trong đặc điểm, cấu tạo, thành phần của các món ăn. Công việc phân loại sẽ trở nên khó khăn không chỉ cho mô hình mà còn khó khăn cho cả con người. Tuy nhiên, mô hình cho kết quả tốt nhất là DenNet121 với 83% là một kết quả tốt và có tiềm năng để phát triển thêm.

6 Kết luận và hướng phát triển

Trong bài báo này, chúng tôi đã xây dựng bộ dữ liệu với các ảnh thu thập từ các trang tìm kiếm phổ biến và thực hiện kiểm tra chất lượng của hình ảnh và cuối cùng thu được 12017 ảnh với 15 món ăn phổ biến ở Việt Nam. Sau đó chúng tôi đã tiến hành áp dụng các kỹ thuật máy học, học sâu để giải quyết bài toán phân loại món ăn mà chúng tôi đã nêu ở mục 1. Với thuật toán máy học truyền thống là VGG-16 + Logistic Regression chúng tôi thu được độ chính xác đạt 71.35% và Top-5 Accuracy là 95%. Và kết quả thu được từ các phương pháp học sâu khá khả quan, DenseNet121 Accuracy 83% và Top-5 Accuracy 98%, MobileNetV2 Accuracy 67% và Top-5 Accuracy 91%.

Bộ dữ liệu về món ăn Việt Nam và các kết quả ban đầu đã cho thấy sự khả quan của bài toán. Với các kết quả đạt được, chúng tôi tin rằng đây sẽ là một trong những nền tảng để phát triển cho việc nhận dạng món ăn truyền thống Việt Nam, nâng cao sự quảng bá văn hóa ẩm thực Việt tới du khách quốc tế qua hình ảnh mà họ chụp được.

Trong tương lai chúng tôi sẽ cải thiện, mở rộng bộ dữ liệu bằng cách thu thập thêm hình ảnh các món ăn từ nhiều nguồn khác nhau một cách bao quát và đảm bảo về chất lượng ảnh. Bên cạnh đó chúng tôi sẽ thử nghiệm thêm một số kiến trúc mạng học sâu khác để cải thiện hiệu suất phân loại món ăn. Chúng tôi hy vọng bộ dữ liệu này sẽ là một trong những bộ dữ liệu tiềm năng để phát triển các phương pháp phân loại món ăn Việt Nam một cách tự động, cũng như góp một phần công sức cho cộng đồng nghiên cứu thị giác máy tính ở Việt Nam.

Tài liệu

1. Faizan Ahmad, Aaima Najam, and Zeeshan Ahmed. Image-based face detection and recognition: " state of the art". *arXiv preprint arXiv:1302.6379*, 2013.

2. Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.
3. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.
4. Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
5. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
6. Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: a dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167*, 2019.
7. Z Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize f1 score. *arXiv preprint ArXiv:1402.1892*, 14, 2014.
8. David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.
9. Simon Mezgec and Barbara Koroušić Seljak. Nutrinet: a deep learning food and drink image recognition system for dietary assessment. *Nutrients*, 9(7):657, 2017.
10. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
11. Azusa Sawada, Eiji Kaneko, and Kazutoshi Sagi. Trade-offs in top-k classification accuracies on losses for deep learning. *arXiv preprint arXiv:2007.15359*, 2020.
12. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
13. Reza Dea Yogaswara and Adhi Dharma Wibawa. Comparison of supervised learning image classification algorithms for food and non-food objects. In *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, pages 317–324. IEEE, 2018.
14. Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu. Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2110–2118, 2016.